Львівський національний університет імені Івана Франка

На правах рукопису

Волощук Орест Романович

УДК 004.853+004.855.5

Методи кластеризації на великих масивах даних

01.05.03 — математичне та програмне забезпечення обчислювальних машин і систем

Магістерська робота

Науковий керівник

Годич Олесь Васильович,

кандидат фізико-математичних наук, доцент

Львів — 2011

ЗМІСТ

ВСТУП

Актуальність теми. Сьогодні все частіше виникають задачі, так чи інакше пов'язані із розбиттям масиву об'єктів на групи за певними критеріями. Із розвитком обчислювальної техніки збільшуються також об'єми баз даних, що можуть піддаватись такому аналізу. Виникає необхідність створення оптимальних алгоритмів обробки великих масивів даних.

Мета і завдання дослідження. Метою дослідження є розвиток методики кластеризації великих масивів даних. Для досягнення цієї мети були сформульовані та вирішені такі основні завдання:

— провести детальний аналіз та дослідити ефективність різноманітних алгоритмів кластеризації;

— здійснити контроль якості результатів роботи алгоритмів;

— розробити методи та виявити можливі оптимізації, що дозволять прискорити виконання задачі кластеризації

Об'єкт дослідження. Об'єктом дослідження є методи ієрархічні та плоскі алгоритми кластеризації даних.

РОЗДІЛ 1
ОГЛЯД СТАНУ ПРОБЛЕМИ ТА ОСНОВНІ ПОНЯТТЯ

Термін „кластерний аналіз", вперше використаний Тріоном у [3], означає набір підходів та алгоритмів, призначених для об'єднання схожих об'єктів у групи. Ця технологія знайшла своє застосування в цілій низці галузей наук та є необхідною частиною більшості сучасних засобів аналізу даних. В 1959 радянський вчений Терентьєв розробив так званий „метод кореляційних плеяд" [4], покликаний здійснювати групування на базі корелюючих ознак об'єктів. Займаючись вивченням кореляцій між різними ознаками озерної жаби, він об'єднав їх в групи за абсолютною величиною коефіцієнту кореляції. Таким чином він отримав дві групи ознак – ознаки із великим та з малим значенням кореляції. Терентьєв назвав ці групи „кореляційними плеядами" та опублікував декілька методів їх аналізу. Це посприяло розвитку методів кластеризації за допомогою графів. На початку 50х років також вийшли публікації Р. Льюїса, Е. Фікса та Дж. Ходжеса, присвячені ієрархічним алгоритмам кластеризації. Відчутний поштовх технології кластерного аналізу дали роботи Розенблатта про розпізнаючий пристій „перцептрон", котрі поклали початок теорії „розпізнавання без вчителя". Поштовхом до розробки методів кластеризації стала публікація [2] в 1963 році. В своїй роботі автори виходили з того, що для створення ефективних біологічних класифікацій процедура кластеризації повинна використовувати всеможливі показники, що характеризують досліджувані організми, проводити оцінку ступеня схожості між цими організмами, та забезпечувати розташування схожих організмів в одну групу. При цьому сформовані групи повинні бути досить локальними, тобто схожість організмів всередині групи повинна бути більшою, ніж між групами. Подальший аналіз

таких груп допоможе вияснити, чи відповідають вони реальним біологічним класифікаціям. Сокел та Сніт, автори цеї роботи, вважали, що виявлення структури розподілу об'єктів у групи допоможе встановити процес утворення цих груп.

Багато дослідників у своїй роботі змушені зіткнутись із необхідністю сформувати змістовні структури із набору спостережень. В біології часто постає задача розділити тварин за видами, користуючись певними вимірюваними ознаками кожної із них. Соціологи використовують техніки кластеризації для виділення соціальних груп за певними ознаками.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *Rosenblatt, F.* The perceptron–a perceiving and recognizing automaton: Tech. Rep. 85-460-1 / F. Rosenblatt: Cornell Aeronautical Laboratory, 1957.
2. *Socal, R.* Principles of Numerical Taxonomy / R. Socal, P. Sneath. — San Francisco: W.H. Freeman, 1963.
3. *Tryon, R. C.* Cluster Analysis / R. C. Tryon. — Ann Arbor: Edwards Brothers, 1939.
4. *Терентьев П. В.* Метод корреляционных плеяд / Терентьев П. В. — Вестник Ленинградского университета, 1959.